Mesure sans contact de la fréquence par caméra basée sur l'apprentissage profond

Yassine Ouzar, Frédéric Bousefsaf, Choubeila Maaoui

Mesure sans contact de la fréquence par caméra basée sur l'apprentissage profond

Yassine Ouzar, Frédéric Bousefsaf et Choubeila Maaoui
Laboratoire de Conception, Optimisation et Modélisation des Systèmes (LCOMS),
Université de Lorraine, LCOMS, F-57000 Metz, France.
{nom.prénom}@univ-lorraine.fr

Résumé

Nous présentons dans ce papier une nouvelle architecture de bout en bout basée sur un réseau de neurones spatio-temporel profond pour l'estimation de la fréquence cardiaque à partir des trames vidéo issues d'une webcam bas coût. Contrairement aux méthodes existantes, nous estimons la valeur de la fréquence cardiaque directement sans passer par l'extraction du signal iPPG et sans incorporer de connaissances préalables ou d'étapes de traitement supplémentaires. Nous avons construit notre réseau en utilisant des couches de convolution séparable en profondeur 3D avec des connexions résiduelles pour extraire simultanément des caractéristiques spatiales et temporelles. Ceci est très approprié pour la mesure en temps réel car le modèle nécessite un nombre réduit de paramètres et un court fragment vidéo. Les résultats obtenus semblent très satisfaisants et prometteurs, d'autant plus que les expériences ont été menées sur des ensembles de données collectés dans des conditions non contrôlées. La mesure de paramètres physiologiques sans contact est à la fois prometteuse et pertinente dans le contexte du suivi de l'évolution de maladies invalidantes. Les avancées récentes sont aujourd'hui intégrées dans des systèmes d'assistance à la personne et sont utilisées durant les séances de thérapie par réalité virtuelle.

Mots-clés : fréquence cardiaque; sans contact; webcam; iPPG; réseaux de neurones convolutifs.

1 Introduction

Selon les dernières statistiques de l'Organisation Mondiale de la Santé, les maladies cardiovasculaires sont la première cause de décès dans le monde (World Health Organisation, 2018). Elles augmentent avec l'augmentation de la population, l'obésité et la sédentarité. Le contrôle non optimal de ces maladies est responsable de 70% des accidents vasculaires cérébraux, de 50% des crises cardiaques et de plusieurs cas d'insuffisance rénale. Ce type de maladies est souvent asymptomatique ce qui nécessite un contrôle périodique et à long terme via des mesures fréquentes de l'activité cardiaque afin de les prévenir et de mieux les prendre en charge.

L'électrocardiographie (ECG) et la photopléthysmographie (PPG) sont les principaux moyens pour la mesure de l'activité cardiaque. Les deux techniques utilisent des capteurs en contact qui doivent être attachés aux parties du corps et nécessitent le respect de certaines conditions pour obtenir de bonnes mesures. Malgré la grande précision et la robustesse fournies par ces dispositifs intrusifs, le contact avec la peau peut être gênant voire infaisable en raison de certains cas critiques citons par exemple les brûlures, les ulcères cutanés, les maladies contagieuses (Sun, 2016). Par conséquent, ces différentes limites, ainsi que la forte demande pour une technologie fiable, confortable, simple, portable, non

stressante et peu coûteuse, ont incité les chercheurs à développer de nouvelles techniques de mesure sans contact des signaux physiologiques.

Au cours de la dernière décennie, de grands progrès ont été réalisés pour l'estimation sans contact des paramètres vitaux tels que la fréquence cardiaque à l'aide de la photopléthysmographie par imagerie (iPPG) pour surmonter les faiblesses des dispositifs invasifs. La iPPG est une technique optique permettant une évaluation à distance de l'activité cardiaque en observant les variations du volume sanguin sur le visage d'une personne à l'aide d'une simple caméra bas coût. Cette technique est très prometteuse en santé publique, en particulier dans le contexte du vieillissement et des maladies invalidantes. Elle est désormais intégrée aux technologies d'assistance (Tagnithammou, 2021).

Les algorithmes d'iPPG classiques sont basés sur des approches conventionnelles qui impliquent généralement des pipelines à plusieurs étages et nécessitent plusieurs étapes de traitement d'image et de signal (Bousefsaf, 2013; de Haan, 2013; Poh, 2010). Ces méthodes ont été mises en œuvre dans des scénarios contraints et reposent sur certaines hypothèses concernant l'interaction lumière-peau et les mouvements de la tête. Par conséquent, la plupart des méthodes proposées fonctionnent raisonnablement bien sur des ensembles de données collectées dans des environnements contrôlés, mais les performances se dégradent considérablement dans des scénarios réels.

Avec le grand succès de l'apprentissage profond pour les tâches d'imagerie médicale et de vision par ordinateur, les travaux récents ont incorporé des architectures d'apprentissage profond à différentes étapes du pipeline de photopléthysmographie conventionnelle (Chen, 2018; Niu, 2020; Yu, 2019). Bien que les méthodes proposées permettent d'extraire avec précision le signal iPPG, mais plusieurs limites restent à surmonter. Tout d'abord, ces systèmes ne sont pas de bout en bout, ce qui nécessite encore des étapes de pré-traitement ou de post-traitement supplémentaire. De plus, la fréquence cardiaque doit être mesurée même dans des scénarios non contrôlés. De nombreuses situations peuvent impacter la mesure : la personne peut bouger la tête ou exprimer des émotions, son visage peut être partiellement occlus ou les conditions d'éclairage peuvent changer en permanence. Cela peut affecter la qualité du signal iPPG extrait et donc dégrader la précision des résultats.

Pour remédier à ces faiblesses, nous avons développé une méthode d'apprentissage profond de bout en bout pour l'estimation instantanée de la fréquence cardiaque directement à partir des séquences vidéo faciales. Notre architecture est entièrement automatique et ne nécessite aucune connaissance préalable ni aucun pré-traitement ou post-traitement particulier. Elle se concentre automatiquement sur les zones les plus vascularisées du visage, analyse les subtiles variations de couleur sur ces régions pour enfin estimer la fréquence cardiaque correspondante.

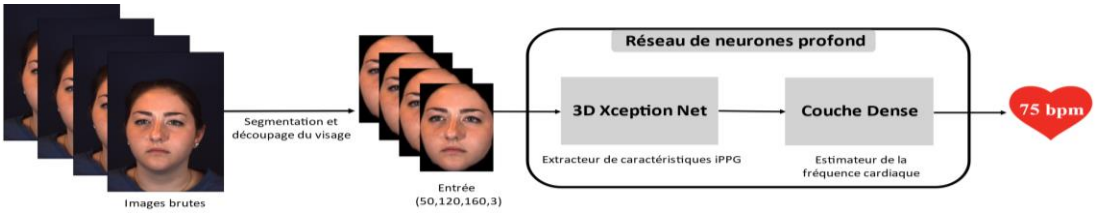

Figure 1 : Aperçu de notre solution proposée pour l'estimation de la fréquence cardiaque instantanée.

2 Matériel et Méthodes

Le framework général de notre méthode est illustré dans la figure 1. Nous considérons la tâche d'estimation de la fréquence cardiaque à partir de vidéos faciales comme une tâche de régression en une étape. Une segmentation du visage est effectuée en premier lieu pour éliminer le fond et les zones non cutanées (Nirkin, 2017). Ensuite, sans aucune étape de prétraitement ou de post-traitement supplémentaire, des lots de 50 images (correspondant à 2 secondes) sont introduits dans un réseau 3D entièrement convolutif pour estimer la fréquence cardiaque correspondante.

2.1 Base de données

Pour fonctionner avec précision dans des scénarios bien contrôlés ainsi que dans des scénarios difficiles, nous avons entraîné notre modèle sur une base de données publique à grande échelle (nommée BP4D+). Cette base de données est dédiée principalement à la reconnaissance multimodale des émotions spontanées à l'aide d'expressions faciales et de paramètres physiologiques tels que la fréquence cardiaque (Zhang, 2016). Par rapport aux bases de données de fréquence cardiaque existantes, BP4D+ est considérablement plus importante en termes de quantité de données et de diversité ethnique (noir, blanc, asiatique, hispanique/latino). Cette base de données peut ainsi fournir un apprentissage plus robuste car elle contient de nombreuses scénarios difficiles tels que des mouvements significatifs de la tête, des expressions faciales et des variations de fréquence cardiaque importantes, ainsi qu'une importante diversité en termes de teint de peau qui n'est pas disponible dans les autres bases de données

2.2 Segmentation du visage

L'extraction des régions d'intérêt (ROI) est la première étape de tous les systèmes de mesure de la fréquence cardiaque par caméra (Niu, 2020; Poh, 2010; Yu, 2019). Elle vise à maximiser le rapport signal/bruit (SNR) en éliminant les régions non cutanées qui ne contiennent aucun changement de couleur associé au rythme cardiaque. À notre connaissance, la plupart des systèmes iPPG existants basés sur l'apprentissage profond ont utilisé soit le visage entier, soit une région du visage sélectionnée grâce à des connaissances empiriques. Plusieurs détecteurs de visages et de repères faciaux ont été utilisés pour localiser la ROI (King, 2009; Viola and Jones, 2004; Zhang, 2016). Cependant, ils échouent souvent lorsque les visages présentent des mouvements de tête importants, des variations de pose, des occlusions ou des expressions faciales. De nombreux autres défis affectent également la capacité d'extraction de la ROI, tels que la couleur de peau, l'éclairage et l'arrière-plan.

Pour surmonter les limitations des algorithmes de détection de visage, nous effectuons une segmentation de visage en utilisant un algorithmes de l'état de l'art proposé initialement pour l'échange de visage (Nirkin, 2017). Cette méthode fonctionne idéalement dans toutes les conditions mentionnées ci-dessus sans manquer aucune image. Les visages sont correctement segmentés des arrière-plans et des occlusions avec une grande précision.

2.3 Architecture

Le réseau proposé est inspiré de l'architecture Xception (Chollet, 2017) qui utilise la convolution séparable en profondeur (CSP) au lieu de la convolution classique. Cette dernière est couteuse en termes de temps de calcul et de besoins en mémoire. L'architecture globale de notre modèle est composée de 36 couches convolutives structurées en 14 modules, tous liés par des raccourcis comme dans les réseaux ResNet à l'exception du premier et du dernier module (figure 2). Le réseau étant très profond, ces connexions résiduelles permettent d'éviter le problème de disparition du gradient. Chaque CSP est suivie d'une normalisation par lots pour stabiliser le processus d'apprentissage et accélérer la convergence, et également une fonction d'activation ReLU pour effectuer une cartographie non linéaire. La sortie de l'extraction des caractéristiques est aplatie et introduite à deux couches denses avec respectivement 1024 et 1 neurones, pour estimer la valeur de la fréquence cardiaque.

2.4 Implémentation

Le modèle proposé est mis en œuvre à l'aide du framework Keras et tensorflow, et exécuté sur NVIDIA Quadro P400. Pour toutes les expériences, l'entrée est fixée à 50 images. Inspiré par la procédure d'optimisation SWATS (Keskar, 2017), nous commençons l'apprentissage avec l'optimiseur Adam rectifié (RAdam) (Liu, 2020), et nous passons à la descente de gradient stochastiques (SGD)

lorsque la précision de l'ensemble de validation cesse de s'améliorer. Nous entrainons le réseau pendant 25 époques avec une taille de lot de 64. Le taux d'apprentissage a été fixé à 10^{-4}. En plus d'une couche dropout d'un ratio de 0,4 appliqué avant la couche dense finale du réseau, des stratégies de régularisation L1 et L2 sont utilisées, ce qui permet de surmonter le problème de surapprentissage et d'améliorer la capacité de généralisation du modèle à de nouvelles données.

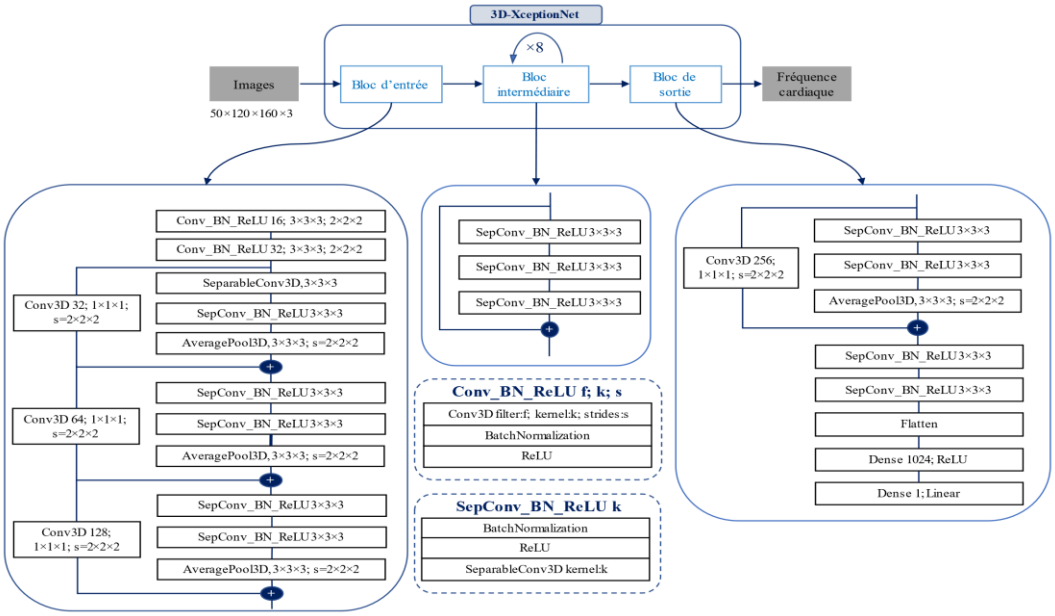

Figure 2 : L'architecture proposée : Elle correspond à une version modifiée du réseau Xception. L'entrée passe d'abord par le flux d'entrée, puis par le flux intermédiaire qui se répète huit fois, et enfin par le flux de sortie qui régresse la valeur de la fréquence cardiaque.

3 Résultats

Afin d'étudier la capacité de généralisation et l'efficacité du modèle proposé présenté, trois bases de données ont été utilisées, à savoir MMSE-HR (Zhang, 2016), MAHNOB-HCI (Soleymani, 2012) et UBFC-RPPG (Bobbia, 2017). MMSE-HR est directement utilisée pour les tests sans aucun traitement supplémentaire car elle a été collectée dans les mêmes conditions que la base d'apprentissage. Alors que UBFC-RPPG et MAHNOB-HCI sont sous-échantillonnés de 30 fps et 61 fps respectivement à 25 fps. Nous évaluons les performances de notre approche avec d'autres techniques de l'état de l'art en utilisant différentes métriques. Les résultats de comparaisons présentés dans les tableaux suivants montrent la grande précision de notre méthode qui surpasse tous les algorithmes de l'état de l'art.

Méthode	**MAE (bpm)**	**RMSE (bpm)**	**r**
PhysNet	12.76	13.25	0.44
SAMC	12.24	11.37	0.71
RhythmNet	6.98	7.33	0.78
AutoHR	5.71	5.87	0.89
Méthode proposée	**4.13**	**5.34**	**0.89**

Tableau 1: Cross-dataset sur MMSE-HR.

Méthode	MAE (bpm)	RMSE (bpm)	r
rPPGNet	5.51	7.82	0.78
SAMC	4.96	6.23	0.83
AutoHR	3.78	5.10	0.86
RhythmNet	-	3.99	0.87
Méthode proposée	**3.17**	**3.93**	**0.88**

Tableau 2 : Résultats sur MAHNOB-HCI.

Méthode	MAE (bpm)	RMSE (bpm)	std
Green	10.2	20.6	20.2
POS	5.12	10.5	10.4
3DCNN	5.45	8.64	8.55
PRNet	5.29	7.24	6.45
Méthode proposée	**4.99**	**6.26**	**6.25**

Tableau 3 : Résultats sur UBFC-RPPG.

4 Conclusion et Perspectives

Dans cet article, nous proposons une nouvelle architecture de bout en bout basée sur un réseau spatio-temporel profond qui prédit la fréquence cardiaque sans passer par l'extraction du signal iPPG et sans utiliser des connaissances préalables. Le réseau proposé s'inspire d'un modèle Xception qui s'est avéré efficace pour les bases de données d'images 2D à usage général en termes de précision, de vitesse de convergence rapide et de faibles coûts de calcul. Nos expériences approfondies ont montré l'efficacité de notre approche qui atteint une plus grande précision et surpasse les méthodes existantes sur trois ensembles de données de référence populaires tels que MMSE-HR, UBFC-RPPG et MAHNOB-HCI. Cependant, nous avons identifié plusieurs problèmes qui peuvent encore être améliorés dans des études futures. Premièrement, les mauvaises performances des techniques d'apprentissage profond pour les échantillons minoritaires dans le cas d'ensembles de données déséquilibrés qui sont fortement biaisés vers une peau plus claire et des fréquences cardiaques moyennes. L'application de stratégies avancées d'augmentation des données ou l'utilisation de données synthétiques pourrait améliorer encore les performances en augmentant le nombre d'échantillons pour les peaux foncées ou pour les fréquences cardiaques faibles et élevées. De plus, nous avons remarqué un taux élevé de valeurs aberrantes et de signaux PPG de mauvaise qualité dans les bases de données que nous avons utilisées. La préparation et le nettoyage des données avant la formation sont essentiels pour entraîner correctement le réseau et éviter les problèmes de surapprentissage. Enfin, les réseaux existants sont souvent constitués d'un grand nombre de paramètres et nécessitent des coûts de calcul élevés, entravant largement son application sur des appareils à faible consommation d'énergie tels que les téléphones portables. Par conséquent, l'étude de modèles de réseau légers peut considérablement améliorer la vitesse et la précision tout en maintenant des performances similaires ou meilleures.

Nos travaux futurs aborderont les problèmes mentionnés ci-dessus pour construire une architecture sophistiquée qui fonctionne avec précision dans des situations réalistes.

References

World Health, O. (2018). global health estimates 2016: death by Cause, Age, Sex, by country and Region, 2000-2016. Geneva.

Bobbia, S., Macwan, R., Benezeth, Y., Mansouri, A., Dubois, J., 2017. Unsupervised skin tissue segmentation for remote photoplethysmography. Pattern Recognition Letters.

Bousefsaf, F., Maaoui, C., Pruski, A., 2013. Continuous wavelet filtering on webcam photoplethysmographic signals to remotely assess the instantaneous heart rate. Biomedical Signal Processing and Control 8, 568–574.

Chen, W., McDuff, D., 2018. Deepphys: Video-based physiological measurement using convolutional attention networks, in: Proceedings of the European Conference on Computer Vision (ECCV). pp. 349–365.

Chollet, F., n.d. Xception: Deep Learning with Depthwise Separable Convolutions 8.

de Haan, G., Jeanne, V., 2013. Robust pulse rate from chrominance-based rPPG. IEEE Transactions on Biomedical Engineering 60, 2878–2886.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., n.d. ImageNet: A Large-Scale Hierarchical Image Database 2.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs].

Keskar, N.S., Socher, R., 2017. Improving Generalization Performance by Switching from Adam to SGD. arXiv:1712.07628 [cs, math].

King, D.E., n.d. Dlib-ml: A Machine Learning Toolkit 4.

Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., Han, J., 2020. On the Variance of the Adaptive Learning Rate and Beyond. arXiv:1908.03265 [cs, stat].

Nirkin, Y., Masi, I., Tran, A.T., Hassner, T., Medioni, G., 2017. On Face Segmentation, Face Swapping, and Face Perception. arXiv:1704.06729 [cs].

Niu, X., Shan, S., Han, H., Chen, X., 2020. RhythmNet: End-to-End Heart Rate Estimation From Face via Spatial-Temporal Representation. IEEE Trans. on Image Process. 29, 2409–2423.

Poh, M.-Z., McDuff, D.J., Picard, R.W., 2010. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. Opt. Express 18, 10762.

Ruder, S., 2017. An overview of gradient descent optimization algorithms. arXiv:1609.04747 [cs].

Soleymani, M., Lichtenauer, J., Pun, T., Pantic, M., 2012. A Multimodal Database for Affect Recognition and Implicit Tagging. IEEE Trans. Affective Comput. 3, 42–55.

Sun, Y., Thakor, N., 2016. Photoplethysmography revisited: from contact to noncontact, from point to imaging. IEEE Transactions on Biomedical Engineering 63, 463–477.

Tagnithammou, T., Monacelli, É., Ferszterowski, A., Trénoras, L., 2021. Emotional state detection on mobility vehicle using camera: Feasibility and evaluation study. Biomedical Signal Processing and Control 66, 102419.

Viola, P., Jones, M., n.d. Rapid Object Detection using a Boosted Cascade of Simple Features 9.

Yu, Z., Li, X., Niu, X., Shi, J., Zhao, G., 2020. AutoHR: A Strong End-to-End Baseline for Remote Heart Rate Measurement With Neural Searching. IEEE Signal Process.

Yu, Z., Li, X., Zhao, G., 2019. Remote Photoplethysmograph Signal Measurement from Facial Videos Using Spatio-Temporal Networks, in: BMVC.

Zhang, K., Zhang, Z., Li, Z., n.d. Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks.

Zhang, Z., Girard, J.M., Wu, Y., Zhang, X., Liu, P., Ciftci, U., Canavan, S., Reale, M., Horowitz, A., Yang, H., Cohn, J.F., Ji, Q., Yin, L., 2016. Multimodal Spontaneous Emotion Corpus for Human Behavior Analysis, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).